D144–D149 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 12 November 2011*
doi:10.1093/nar/gkr965

AnimalTFDB: a comprehensive animal transcription factor database

Hong-Mei Zhang[1], Hu Chen[1], Wei Liu[1], Hui Liu[1], Jing Gong[1], Huili Wang[2] and An-Yuan Guo[1,*]

[1]Hubei Bioinformatics & Molecular Imaging Key Laboratory, Department of Systems Biology, College of Life Science and Technology and [2]Department of Environment Engineering, Huazhong University of Science and Technology Wenhua College, Wuhan, 430074, China

Received August 6, 2011; Revised September 12, 2011; Accepted October 15, 2011

ABSTRACT

Transcription factors (TFs) are proteins that bind to specific DNA sequences, thereby playing crucial roles in gene-expression regulation through controlling the transcription of genetic information from DNA to RNA. Transcription cofactors and chromatin remodeling factors are also essential in the gene transcriptional regulation. Identifying and annotating all the TFs are primary and crucial steps for illustrating their functions and understanding the transcriptional regulation. In this study, based on manual literature reviews, we collected and curated 72 TF families for animals, which is currently the most complete list of TF families in animals. Then, we systematically characterized all the TFs in 50 animal species and constructed a comprehensive animal TF database, AnimalTFDB. To better serve the community, we provided detailed annotations for each TF, including basic information, gene structure, functional domain, 3D structure hit, Gene Ontology, pathway, protein–protein interaction, paralogs, orthologs, potential TF-binding sites and targets. In addition, we collected and annotated transcription cofactors and chromatin remodeling factors. AnimalTFDB has a user-friendly web interface with multiple browse and search functions, as well as data downloading. It is freely available at http://www.bioguo.org/AnimalTFDB/.

INTRODUCTION

Regulation of gene expression controls the spatial and temporal expression pattern and influences all biological processes in organisms. In this regulation, transcriptional regulatory system plays a key role and involves diverse proteins, including RNA polymerase, basal and sequence specific DNA-binding transcription factors (TFs), transcription cofactors and chromatin remodeling proteins (1). Among them, TFs are most fascinating owing to their complex regulation function. Here we use the common definition of TFs, which are proteins containing a sequence specific DNA-binding domain (DBD) and regulating target gene transcription. Based on their DBDs, TFs could be classified into different TF families. It is reported that about half of the TF families in plants and animals are plant or animal specific (2). TF families in plants were well characterized and several databases for plant TFs were developed (3–5). However, until now, there is no a comprehensive animal TF family list and a database characterizing all the TFs based on TF families for the sequenced animal genomes.

To date, there are several databases about TFs for some animals, such as TFdb for mouse (6), FlyTF for fruit fly (7), TFCat for human and mouse (8), TFCONES for human, mouse and fugu (9) and ITFP for human, mouse and rat (10). As mentioned, these databases only focus on single or a few genomes. Although TRANSFAC collects abundant information about TFs for several kinds of animals (11), yet it is a commercial database and collected only experimentally verified TFs. DBD is a comprehensive TF database for more than 900 genomes across the three super kingdoms of life (Bacteria, Archaea and Eukaryotes) and includes dozens of animals (12). However, the TF family classification and TF annotation for animals could be improved to better serve the community. Thus, an integrated animal TF database with higher coverage, higher accuracy and full annotation is required as more and more animal genomes were sequenced.

With this in mind, we collected and curated a comprehensive list for animal TF families by manual literature reviews. Then we predicted TFs for all these families in 50 sequenced animal genomes and constructed a comprehensive animal TF database AnimalTFDB (http://www.bioguo.org/AnimalTFDB/). Moreover, we predicted transcription cofactors and chromatin remodeling

*To whom correspondence should be addressed. Tel: +86 27 8779 3177; Fax: +86 27 8779 3177; Email: guoay@mail.hust.edu.cn

factors for these 50 genomes. The database has a user-friendly interface to display and search the detailed annotations. We hope that AnimalTFDB may become a useful resource for the research community, especially in the studies of comparative genomics and transcriptional regulation.

METHODS

Data sources

Currently, AnimalTFDB contains TFs, transcription cofactors and chromatin remodeling factors identified in 50 animals (Table1). All genome data were downloaded from Ensembl (release version 60, http://www.ensembl.org/) database.

Animal TF family list and their HMM profiles

We characterized and classified TFs by their sequence specific DBDs. After reviewing literatures, we finally collected and curated 71 animal TF families and a group named 'others' including some orphan TFs (http://www.bioguo.org/AnimalTFDB/help.php), which is currently the most complete TF family list for animals. Among them, 59 families had Hidden Markov Model (HMM) profiles for their DBDs in Pfam database (v25.0) (13), while no HMM profiles were available for the other 12 TF families. We built HMM profiles for them based on their DBD multiple sequence alignments by the hmmbuild program in the HMMER package.

TFs identification

We applied the hmmsearch program in HMMER package to search all the protein sequences against the DBD HMM profiles to predict TFs. Based on our manual checking for the predicted human and mouse TF results, we took *E*-value 0.0001 as the cutoff, which simultaneously considered the accuracy and sensitivity. For TFs that had more than one DBD, we assigned them into families based on their true DBD, which is the domain exactly binding to DNA in those proteins.

Identification of transcription cofactors and chromatin remodeling factors

In AnimalTFDB, transcription cofactors were considered as proteins that interact with TFs in the transcription apparatus but are not able to bind the DNA directly. The chromatin remodeling factors were defined as proteins that regulate transcription by modifying the chromatin formation. To identify them, we firstly got the human transcription cofactors and chromatin remodeling factors from TFCONES (9) and Gene Ontology (GO) (14) databases according to the GO items: transcription cofactor activity and chromatin remodeling, respectively. Then, we used the human sequences to perform BLAST search and chose the best BLAST hits as the transcription cofactors or chromatin remodeling factors for the searched species.

DATABASE CONTENT

Annotations of the identified factors

The numbers of TFs, transcription cofactors and chromatin remodeling factors identified in 50 animals were showed in Table 1. In order to provide more useful information, we made extensive annotations for them. We obtained the basic gene information and GO annotation from NCBI and Ensembl databases. Putative functional domains and 3D structure hits for the longest protein of each gene were offered. The protein–protein interaction information was parsed from BioGRID (15), HPRD (16) and An atlas of human and mouse TF interactions (17) databases. The pathway annotations from BioCarta (http://www.biocarta.com/) and KEGG (18) databases were available in AnimalTFDB. TFs binding sites and target genes were extracted from TRED (19) and JASPAR (20) databases. In addition, we also provided links to GenBank, Unigene and many species-specific databases such as: MGI, HGNC, FlyBase and so on.

Putative ortholog and paralog annotation

To predict the putative orthologs of these factors among different species, the reciprocal best hit (RBH) method (21) was used. We performed the all-against-all BLASTP search between proteins of two genomes with strict cutoffs *E*-value $\leq 1e{-}20$, coverage $\geq 70\%$, identity $\geq 50\%$ and set the reciprocal best hit pairs as orthologs. While, we applied the BLAST score ratio (BSR) (22) approach to predict paralogs. BLASTP search was done in each genome with the same benchmark applied in ortholog finding. After comparing the results of different BSR value, we chose the BSR value 0.4 as the cutoff for paralogs.

WEB INTERFACE

Database organization

Considering MySQL is a free database management system widely applied in bioinformatics, we stored all the information of AnimalTFDB in a MySQL database. Since the different TF annotations varied in contents and formats, we classified all the data into 30 separated tables. The Ensembl ID and Gene ID were used as the main keys to organize and link all the tables.

Data browse

To help users browse the data conveniently and clearly, AnimalTFDB provided two different ways to browse the data: (i) browse by species; (ii) browse by family. On the browse family page, all TF families were further merged into six groups based on the TRANSFAC classification: helix–turn–helix, other α-helix, zinc-coordinating, basic domains, β-scaffold and unclassified structure. The TF family list in each group was shown by the treeview on the left part of this page and the 3D structure images of TF DBDs were used as the family logos on the right part. On the browse species page, 50 species were classified into

Table 1. Numbers of TFs, transcription cofactors and chromatin remodeling factors of 50 species in current AnimalTFDB

Group names	Species	Common names	TFs	CoFs	CRFs	Total
Primates	*Homo sapiens*	Human	1544	302	150	1996
	Macaca mulatta	Macaque	1440	266	119	1825
	Pan troglodytes	Chimpanzee	1429	272	135	1836
	Gorilla gorilla	Gorilla	1429	264	130	1823
	Callithrix jacchus	Marmoset	1397	277	132	1806
	Pongo pygmaeus	Orangutan	1331	263	118	1712
	Microcebus murinus	Mouse Lemur	1037	180	77	1294
	Otolemur garnettii	Bushbaby	894	129	72	1095
	Tarsius syrichta	Tarsier	842	151	64	1057
Rodents	*Mus musculus*	Mouse	1457	279	130	1866
	Rattus norvegicus	Rat	1371	257	119	1747
	Cavia porcellus	Guinea Pig	1054	253	117	1424
	Oryctolagus cuniculus	Rabbit	1047	252	117	1416
	Ochotona princeps	Pika	903	173	75	1151
	Dipodomys ordii	Kangaroo rat	862	170	78	1110
	Tupaia belangeri	Tree Shrew	815	138	64	1017
	Spermophilus tridecemlineatus	Squirrel	810	128	52	990
Laurasiatheria	*Bos taurus*	Cow	1313	257	123	1693
	Equus caballus	Horse	1240	258	123	1621
	Ailuropoda melanoleuca	Giant Panda	1199	258	127	1584
	Tursiops truncatus	Dolphin	1167	234	110	1511
	Pteropus vampyrus	Megabat	1119	236	111	1466
	Canis familiaris	Dog	1062	257	129	1448
	Sus scrofa	Pig	1038	195	90	1323
	Myotis lucifugus	Microbat	970	156	69	1195
	Felis catus	Cat	887	139	62	1088
	Erinaceus europaeus	Hedgehog	744	118	63	925
	Sorex araneus	Shrew	630	126	61	817
	Vicugna pacos	Alpaca	646	118	58	822
Afrotheria	*Loxodonta africana*	Elephant	1096	261	119	1476
	Procavia capensis	Hyrax	983	177	74	1234
	Echinops telfairi	Lesser hedgehog tenrec	985	155	59	1199
Xenarthra	*Dasypus novemcinctus*	Armadillo	868	132	61	1061
	Choloepus hoffmanni	Sloth	725	107	48	880
Other mammals	*Monodelphis domestica*	Opossum	1454	241	97	1792
	Macropus eugeni	Wallaby	897	150	53	1100
	Ornithorhynchus anatinus	Platypus	814	149	60	1023
Birds and reptiles	*Taeniopygia guttata*	Zebra Finch	1185	181	82	1448
	Gallus gallus	Chicken	775	192	83	1050
	Anolis carolinensis	Lizard	1211	197	82	1490
Amphibia	*Xenopus tropicalis*	Frog	1038	168	67	1273
Fishes	*Danio rerio*	Zebrafish	1916	160	77	2153
	Takifugu rubripes	Fugu	1274	162	73	1509
	Tetraodon nigroviridis	Tetraodon	1292	151	63	1506
	Gasterosteus aculeatus	Stickleback	1227	153	69	1449
	Oryzias latipes	Medaka	1187	138	63	1388
Other chordates	*Ciona savignyi*	Sea squirt	409	32	19	460
	Ciona intestinalis	Sea squirt	428	40	16	484
Other Eukaryotes	*Drosophila melanogaster*	Fruitfly	627	38	20	685
	Caenorhabditis elegans	Worm	657	21	9	687
Total			52 725	9111	4169	66 005

CoFs, transcription cofactors; CRFs, chromatin remodeling factors.

11 categories according to the Ensembl taxonomy, which were primates, rodents, laurasiatheria, afrotheria, xenarthra, other mammals, birds & reptiles, amphibians, fishes, other chordates and other eukaryotes. An image from Ensembl was used to show phylogenetics of the 50 animals and an equivalent treeview was built on the left part. Users can browse data by clicking the logos of family and species or by clicking the name on the left treeview of the browse pages. In AnimalTFDB, a cascading style is applied for data browsing, which is browsed by the steps species->families->family gene list->single gene annotation or families->species->family gene list->single gene annotation (Figure 1).

Data search

AnimalTFDB provided two different ways to search the data: quick search and advanced search. A quick search box was shown at the top-right of each page designed for searching by Ensembl IDs for gene, transcript and protein, Entrez gene ID or gene symbol. Advanced search page provided multiple ways for searching by different

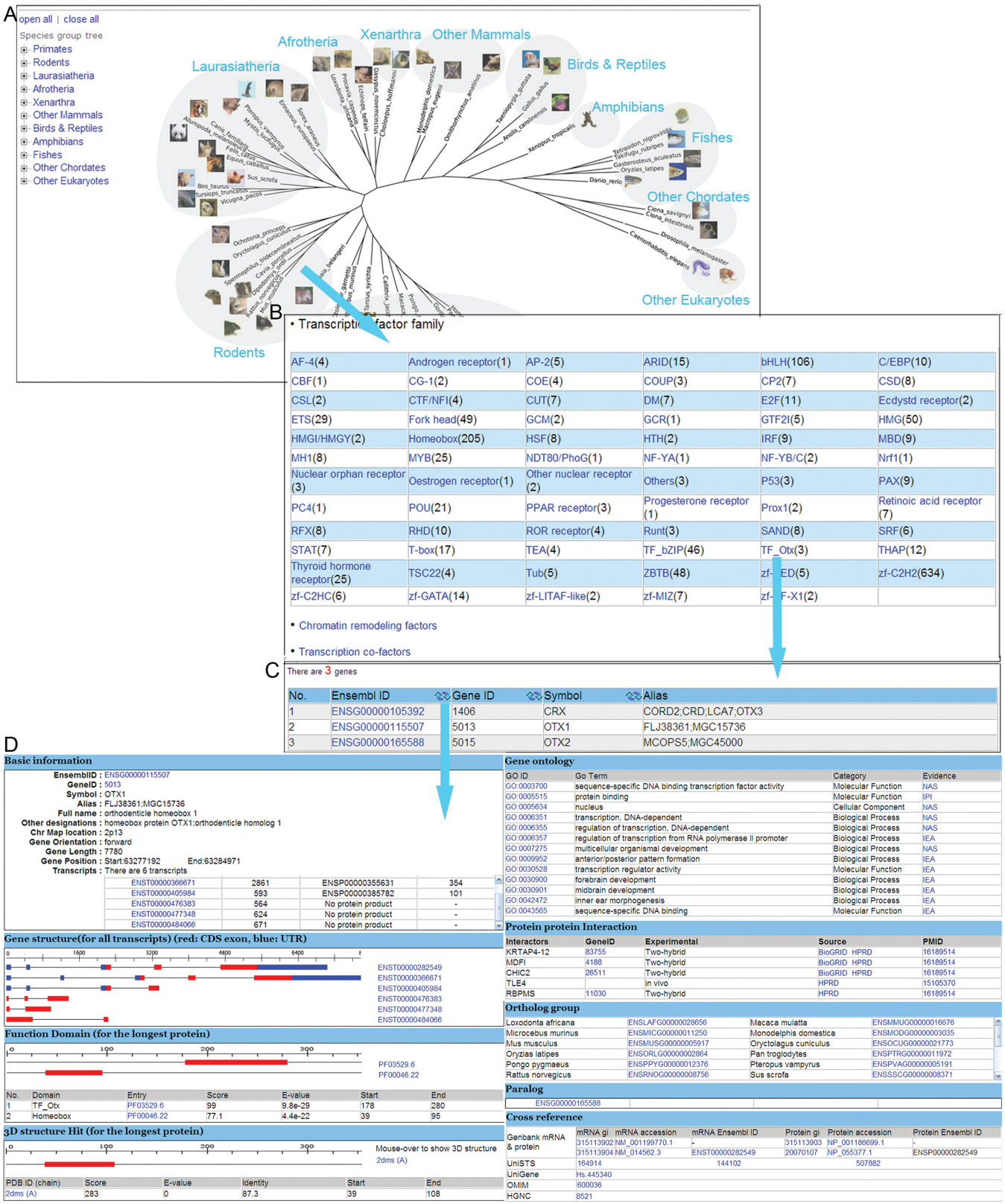

Figure 1. An overview and gene annotation page in AnimalTFDB. (**A**) Species in AnimalTFDB. (**B**) Three kinds of factors in human: TFs, transcription cofactors and chromatin remodeling factors. (**C**) A list of human TFs in the TF_Otx family. (**D**) An example of gene annotation page.

annotations and keywords of each gene. In addition, users could assign the specific families and species for better search.

DISCUSSION

Comparison with other databases and evaluation of TF identification

We compared our predicted human and mouse TFs with those published by DBD (12) and TFCat (8) databases. DBD is a comprehensive predicted TF database for bacteria, archaea and eukaryotes, while TFCat is a curated catalog for human and mouse TFs. For DBD database, through converting the protein ID into gene ID, we obtained 1383 and 1386 Ensembl gene IDs for human and mouse TF genes, respectively. By comparison, the AnimalTFDB includes 93.7% of human TFs and 93.6% of mouse TFs from DBD database. For the TFs in TFCat database, after ID conversion, we got 521 and 543 Ensembl gene IDs for human and mouse TFs, respectively. The compared result showed that 97.1% of human TFs and 96.3% of mouse TFs from TFCat database were available in our database.

We carefully checked the difference between our AnimalTFDB with the two other databases. For those TFs in the two databases but not in our database, there are two cases. First, some of them are not true TFs predicted by false TF DBD models, such as zf-A20, RNA_pol_Rpb2 and SART-1. Second, some of them should be transcription cofactors or chromatin remodeling factors, which are in the corresponding lists of AnimalTFDB. We also examined the approximately 300 AnimalTFDB-specific TFs for human and mouse. The results showed that some of them were predicted by our unique TF families, such as THAP, CBF, TSC22, Nrf1 and COE. Proteins in these families are true TFs evidenced by publications or having a typical DBD. About half of AnimalTFDB specific TFs were distributed in zf-C2H2, Homeobox, HMG and MYB families, which are all big TF families and account for ~60% TFs of the genome. Although most of the specific TFs in these big families are unknown proteins containing typical DBDs, we still found a few of them (e.g. KLF6, KLF8, PBX2, TCF7L1 and HBP1) are proved to be as TFs by experiments in publications. Thus, we think we should keep them in the database.

Furthermore, we used the GO annotations to evaluate the reliability and accuracy of our TF list. As a result, we found that 96.3% of our identified human TFs were annotated by TF-related GO terms, such as 'TF activity', 'transcription activator/repressor/regulator activity' and 'DNA binding'. These results suggest that the TF prediction approach we used has a reliable performance.

Comparing to other databases, our AnimalTFDB have a more complete and accurate TF family list, and thus a more accurate TF gene list with higher sensitivity and specificity. Moreover, our website is intuitive and easy to browse and search for users. Thirdly, comprehensive annotations are provided in our database as described above. Therefore, we think the AnimalTFDB database will be helpful for the community.

FUTURE PERSPECTIVES

AnimalTFDB is a comprehensive animal TF database, which characterized genome-wide TFs, transcription cofactors and chromatin remodeling factors in 50 sequenced animal genomes. According to their DBDs, all the TFs were classified into 72 families, and this is currently the most complete animal TF family list. Since our pipeline for TF prediction is built, it is much easier for us to update the data regularly with more animal genome data available. Further, we will pay more attention to the transcriptional cofactors and chromatin remodeling factors and try to classify them into different families in the future. We plan to construct and maintain a comprehensive animal TF database to provide a solid foundation for the studies of transcriptional regulation and comparative genomics.

AVAILABILITY

The AnimalTFDB database is freely available at http://www.bioguo.org/AnimalTFDB/.

ACKNOWLEDGEMENTS

The authors would like to thank Zhaowu Ma, Huashan Ye, Mi Zhou, Jun Yan, Shuzhen Kuang, Yifang Liao and Yuliang Wu for their valuable advices to improve the database.

FUNDING

Starting Fund from Huazhong University of Science and Technology (to A.Y.G.); Fundamental Research Funds for the Central Universities (2010MS045); and National Natural Science Foundation of China (31171271). Funding for open access charge: National Natural Science Foundation of China (31171271).

Conflict of interest statement. None declared.

REFERENCES

1. Lemon,B. and Tjian,R. (2000) Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev.*, **14**, 2551–2569.
2. Riechmann,J.L., Heard,J., Martin,G., Reuber,L., Jiang,C., Keddie,J., Adam,L., Pineda,O., Ratcliffe,O.J., Samaha,R.R. *et al.* (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.
3. Guo,A., He,K., Liu,D., Bai,S., Gu,X., Wei,L. and Luo,J. (2005) DATF: a database of Arabidopsis transcription factors. *Bioinformatics*, **21**, 2568–2569.
4. Riano-Pachon,D.M., Ruzicic,S., Dreyer,I. and Mueller-Roeber,B. (2007) PlnTFDB: an integrative plant transcription factor database. *BMC Bioinformatics*, **8**, 42.
5. Guo,A.Y., Chen,X., Gao,G., Zhang,H., Zhu,Q.H., Liu,X.C., Zhong,Y.F., Gu,X., He,K. and Luo,J. (2008) PlantTFDB: a comprehensive plant transcription factor database. *Nucleic Acids Res.*, **36**, D966–D969.

6. Kanamori,M., Konno,H., Osato,N., Kawai,J., Hayashizaki,Y. and Suzuki,H. (2004) A genome-wide and nonredundant mouse transcription factor database. *Biochem. Biophys. Res. Commun.*, **322**, 787–793.
7. Adryan,B. and Teichmann,S.A. (2006) FlyTF: a systematic review of site-specific transcription factors in the fruit fly Drosophila melanogaster. *Bioinformatics*, **22**, 1532–1533.
8. Fulton,D.L., Sundararajan,S., Badis,G., Hughes,T.R., Wasserman,W.W., Roach,J.C. and Sladek,R. (2009) TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol.*, **10**, R29.
9. Lee,A.P., Yang,Y., Brenner,S. and Venkatesh,B. (2007) TFCONES: a database of vertebrate transcription factor-encoding genes and their associated conserved noncoding elements. *BMC Genomics*, **8**, 441.
10. Zheng,G., Tu,K., Yang,Q., Xiong,Y., Wei,C., Xie,L., Zhu,Y. and Li,Y. (2008) ITFP: an integrated platform of mammalian transcription factors. *Bioinformatics*, **24**, 2416–2417.
11. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
12. Kummerfeld,S.K. and Teichmann,S.A. (2006) DBD: a transcription factor prediction database. *Nucleic Acids Res.*, **34**, D74–D81.
13. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
14. Barrell,D., Dimmer,E., Huntley,R.P., Binns,D., O'Donovan,C. and Apweiler,R. (2009) The GOA database in 2009–an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.*, **37**, D396–D403.
15. Stark,C., Breitkreutz,B.J., Chatr-Aryamontri,A., Boucher,L., Oughtred,R., Livstone,M.S., Nixon,J., Van Auken,K., Wang,X., Shi,X. *et al.* (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.*, **39**, D698–D704.
16. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
17. Ravasi,T., Suzuki,H., Cannistraci,C.V., Katayama,S., Bajic,V.B., Tan,K., Akalin,A., Schmeier,S., Kanamori-Katayama,M., Bertin,N. *et al.* (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, **140**, 744–752.
18. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
19. Zhao,F., Xuan,Z., Liu,L. and Zhang,M.Q. (2005) TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies. *Nucleic Acids Res.*, **33**, D103–D107.
20. Portales-Casamar,E., Thongjuea,S., Kwon,A.T., Arenillas,D., Zhao,X., Valen,E., Yusuf,D., Lenhard,B., Wasserman,W.W. and Sandelin,A. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D110.
21. Moreno-Hagelsieb,G. and Latimer,K. (2008) Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics*, **24**, 319–324.
22. Rasko,D.A., Myers,G.S. and Ravel,J. (2005) Visualization of comparative genomic analyses by BLAST score ratio. *BMC Bioinformatics*, **6**, 2.